Електронските ресурси и филолошките студии,

Филолошки факултет "Блаже Конески",

Скопје, 2013, pp. 12-26.

ЗВОНКО НИКОДИНОВСКИ

Универзитет „Св. Кирил и Методиј“, Скопје

ЕЛЕКТРОНСКИТЕ АЛАТКИ НА ЛИНГВИСТОТ

ABSTRACT : À l'ère numérique, le linguiste ne peut pas se passer des outils électroniques. Nous postulons une classification de ces outils en deux catégories : 1. Outils électroniques seravnt à collecter des données linguistiques et 2. Outils électroniques servant à collecter des données métalinguistiques (ou données sur les données linguistiques). Dans la première catégorie, nous passons en revue les navigateurs, les moteurs de recherche et les métamoteurs, les corpus linguistiques, les dictionnaires électroniques et les dictionnaires visuels ainsi que les moteurs de recherche de bureau. Dans la deuxième catégorie, nous établissons de nouveau une division en deux classes : 1. Outils électroniques servant à collecter des données métalinguistiques générales et 2. Outils linguistiques servant à collecter des données métalinguistiques particulières. Au bout de cette analyse, nous concluons qu'une recherche linguistique électronique possède, par rapport à la recherche linguistique classique, certains atouts indéniables : elle est plus rapide, plus exhaustive, plus précise, plus vérifiable, moins chère et plus accessible.

Mots-clés : recherche linguistique, données linguistiques et métalinguistiques, outils électroniques (navigateurs, moteurs de recherche, corpus linguistiques, dictionnaires électroniques, archives linguistiques)

Лингвистиката не престанува да го проучува говорот и јазиците кои се зборуваат/зборувале на сите четири страни на светот. Она што го разликува денешниот лингвист во однос на лингвистот од времето пред појавата на компјутерите и особено на интернетот е употребата на електронскиот дигитален запис. Можноста да се запишат во електронски дигитален код текстуалните, звучните и визуелните податоци претставува значајна алка во развојниот пат на комуникацијата меѓу луѓето која може да се спореди со појавата на писмото или пак со појавата на печатницата. Информациските и комуникациските технологии имаат големо влијание врз сите сфери од човековото живеење, а ние ќе се задржиме само на нивната улога во лингвистичките истражувања.

Во нашиот труд ќе се обидеме да ги опфатиме главните електронски алатки коишто му стојат на располагање на лингвистот. Основните наши поставки во врска со улогата на електронските алатки во лингвистичките истражувања се следните: 1. електронски-потпомогнатото лингвистичко истражување (понатаму: електронското лингвистичко истражување) е **побрзо** 2. електронското лингвистичко истражување е **поопфатно** 3. електронското лингвистичко истражување е **попрецизно** 4. електронското

лингвистичко истражување е **попроверливо** 5. електронското лингвистичко истражување е **поевтино** и 6. електронското лингвистичко истражување е **попристапно**.

Вториот термин од нашиот наслов "алатка" е двозначен и соодветствува на развојот на компјутерите. Ако компјутерот е алатка која има за намена автоматски да пресметува, таа своја функција ја извршува со помош на неопходните два составни дела: хардверот и софтверот. И лингвистот не може да врши електронско истражување без хардвер и без софтвер.

Основниот хардвер без кој нема електронско лингвистичко истражување зависи од потребите на лингвистот, лични или институциски, но приближно би можеле да го претставиме на следниот начин: мултимедијален персонален компјутер со пристап до брз интернет, со доволно брз повеќејадрен процесор со повеќе периферни уреди: мемориски уреди (РАМ меморија, дисковни единици (механички диск, електронски диск, USB меморија, мемориски картички), влезни периферни единици (тастатура, глувче, скенер, микрофон, веб камера, телевизиска картичка/земјена или сателитска/) и излезни единици (графичка карта, екран, звучна карта, звучници, печатач, модем, рутер). Главната компонента на компјутерот, микропроцесорот, има најважна улога во брзината на компјутерот, но и секој друг помошен уред има дополнителна улога во побрзата обработка на разните видови податоци. Основните познавања за хардверскиот дел на компјутерот се потребни за подобро конфигурирање на компјутерот и за негово поефикасно искористување. Доколку се работи за некои поспецифични лингвистички истражувања (фонетски, дијалектолошки и други) потребни се, се разбира, и други специјализирани алатки.

Второто значење на зборот "алатка" нè упатува на софтверскиот дел на компјутерот преку кој всушност и се материјализира улогата на помошно средство во лингвистичките истражувања.

Лингвистот работи со јазичните и говорните форми кои се јавуваат во пишана или звучна форма. Значи, сè она што претставува јазична или говорна единица и што се сретнува во електронска форма (на интернет, на мемориски носачи /разни видови дискови; ЦД-а, ДВД-а, УСБ мемории, разни тврди дискови/, телевизија) може да му користи на лингвистот. Се разбира тоа се сирови податоци што лингвистот треба да ги обработи во своето проучување. Денес, интернетот му овозможува на лингвистот да има достап до неизмерен број на податоци кои можат да му служат во неговите истражувања. Во основа, сите тие податоци можеме да ги групираме во две категории: 1. податоци од или на јазикот и 2. податоци за јазикот. И двата вида се опишуваат со една единствена придавка во западноевропските јазици: анг. "linguistic", додека во македонскиот јазик за првиот вид на податоци го користиме терминот "јазични", додека за вториот вид податоци постои терминот "лингвистички". Оваа поделба ќе ја користиме како основа за претставување на електронските алатки на лингвистот. Треба да

нагласиме дека честопати и двата вида на податоци можат да се најдат на исто место, доколку се работи за структурирани и обработени јазични податоци како што ќе видиме подолу. Би сакале да наведеме уште едно објаснување: во нашето излагање се осврнуваме на алатките кои се користат во оперативниот систем "Windows", иако во најголем дел сите тие алатки се повеќеплатформски и се сретнуваат и во други оперативни системи.

I. ЕЛЕКТРОНСКИ АЛАТКИ ЗА СОБИРАЊЕ ЈАЗИЧНИ ПОДАТОЦИ

Сметаме дека собирањето податоци претставува предуслов за секое научно истражување, зашто истражувачот ги проучува податоците во врска со некоја појава за да утврди одредени ставови или тези со кои се опишува и/или објаснува дадената појава.

Основни алатки преку кои се обавува собирањето јазични податоци од интернет претставуваат најпрвин **прегледувачите** (анг. browsers). Меѓу бројните прегледувачи кои постојат се издвојуваат неколку: Internet Explorer (<http://windows.microsoft.com/en-us/internet-explorer/ie-10-worldwide-languages>), Mozilla Firefox (<http://www.mozilla.org/en-US/firefox/all/>), Chrome (<https://www.google.com/intl/en/chrome/browser/>), Opera (http://-en.softonic.com/s/free-download-opera/english>), Safari (<http://support.apple-.com/kb/dl1531>). Се разбира, секој од нив има свои подобри и полоши страни така што препорачливо е самиот лингвист да испроба повеќе прегледувачи за да може да избере еден или два што ќе ги користи секојдневно. Кога би можело да се искомбинира еден кој ќе ги содржи најдобрите карактеристики на секој од наведените прегледувачи како и од некои други: Maxthon (<http://maxthon.en.softonic.com/>), Avant Browser (<http://www.-avantbrowser.com/download.aspx>), GreenBrowser (<http://greenbrowser.-en.softonic.com/>) и др. би било најдобро, но, за жал, тоа не е можно. Како и да е, денес прегледувачите си личат поприлично еден на друг затоа што се натпреваруваат да вклучат одредени карактеристики кои се среќаваат кај други прегледувачи.

Она што е многу важно за прегледувачите се дополнителните програми (addons, plug-ins) кои ги подобруваат нивните функции и честопати додаваат одредени икони или ленти во самиот прегледувач. Потребно е значи да се инсталираат одредени екстензии за да може подобро да се прегледуваат одредени видови на фајлови (pdf, flash, видео, звучни, e-book фајлови и др.) Сите наведени прегледувачи може да се сретнат во различни јазични верзии а кај некои од нив постои и македонска верзија.

Прегледувачот сам по себе овозможува да се дојде до одредена веб страница доколку се знае нејзината адреса. Изнаоѓањето на страници со нивните адреси е токму задача на **пребарувачите** (анг. search engines) кои се неизбежна алатка за пронаоѓање на информации па според тоа и на

јазични податоци неопходни за лингвистичките истражувања. Денес, најпознат и најмногу користен пребарувач е Google (<https://www.google.-com>. Секако постојат и други пребарувачи меѓу кои Bing (<http://www.-bing.com/>), Yahoo (<http://www.yahoo.com/>), Ask.com (<http://www.ask.-com/>), Lycos (<http://www.lycos.com/>). На теренот на источна Азија доста се користи и кинескиот пребарувач Baidu (<http://www.baidu.com/>). Една општа забелешка за сите пребарувачи е дека бројот на страници кои се пронајдени од страна на пребарувачот е само виртуелна бројка (при некои пребарувања резултатот е бројка од неколку милијарди [1]). Најголемиот број на страници што ги покажува и коишто може да се отворат е околу 800, а во најголем број на случаи се движи на нешто повеќе од 500. Сите пребарувачи користат одредена синтакса, одреден број на оператори, регуларни изрази а имаат и други особини (да пребаруваат разни формати на фајлови, да пребаруваат само во одреден дел од страницата, да пребаруваат само одреден сајт или пак страници кои упатуваат на тој сајт, да пребаруваат страници во ограничен временски период и др.). Кога ќе се пронајде страницата која го содржи бараниот израз, тогаш во самиот прегледувач постои можност да се обележат со посебна боја сите појавувања на тој израз во даден текст. На тој начин можеме да видиме во колкав број и во кои контексти се јавува дадениот израз.

Треба да се спомене дека постојат и програми кои нудат истовремено пребарување преку повеќе пребарувачи. Таквите програми се нарекуваат **метапребарувачи**: Metacrawler (<http://www.metacrawler.com/>), Webcrawler (<http://www.webcrawler.com/>) и др.

Во македонската средина интересни се сервисите на **медиските агрегатори** (анг. **media aggregators)** Daily.mk (<http://daily.mk/>) и Time.-mk (<http://www.time.mk/>) овој последниов содржи и англиско-македонско-англиски речник и архива) кои овозможуваат пребарување на интернетските страници на голем дел од македонските печатени и електронски медиуми.

Треба да се спомене и корисниот сервис Google Ngram Viewer (<http://books.google.com/ngrams>) кој овозможува да се пребарува фреквенцијата на појавување на одреден збор или група зборови во електронските текстови на Google Books и притоа да се избере временскиот период. Ова е корисна можност за историски лексиколошки истражувања, особено што може да се добијат графикони кои содржат повеќе барани ајтеми одеднаш.

Главното прашање што треба да си го постави лингвистот при некое истражување, покрај теорискиот лингвистички модел се разбира, е пред-

[1] Најголем број на страници индексирани од страна на пребарувачот Гугл упатуваат на ознаката “http”, вкупно 25.270.000.000.

метот на истражувањето и корпусот што ќе го искористи за тоа истражување. Откако ќе го постави предметот и ќе го избере теорискиот модел, лингвистот мора да размислува како да го оформи својот корпус. Денеска постојат голем број на **електронски корпуси** [2] (анг. **electronic corpora**) кои во називот го носат името на земјата од каде потекнуваат (Американски национален корпус (<http://www.americannationalcorpus.org/>), Британски национален корпус (<http://www.natcorp.ox.ac.uk/>), Чешки национален корпус (<https://korpus.cz/english/index.php>), Руски национален корпус (<http://ruscorpora.ru/search-paper.html>) [3] и др.). Изработката на електронски корпуси и нивното искористување е дел од корпусната лингвистика во рамките на информатичката лингвистика [4]. Во најголем број се работи за корпуси на пишани текстови но во последните години постојат сè повеќе и корпуси на говорни текстови [5]. Овде може да ги споменеме следните корпуси: Оксфордскиот корпус на англискиот јазик кој е наплатлив за користење и кој содржи околу 2 милијарди токени. Корпусот на францускиот јазик под наслов Wortschatz (Вокабулар) (<http://wortschatz.uni-leipzig.de/-ws_fra/>) на Универзитетот во Лајпциг е прекрасен отворен корпус кој содржи околу 370 милиони токени со над 37 милиони реченици. Добиените резултати ги прикажуваат зборовите кои се наоѓаат во близина на барaниот збор или пак се наоѓаат веднаш лево или десно од бараниот збор. Резултатите на примерите се искажани во целосни реченици во кои бараниот

---

[2] Меѓу многуте страници кои содржат адреси на електронски корпуси за разни јазици можеме да ги наведеме следните: страницата на Европското здружение за јазични ресурси ELRA (European Language Resources Association) <http://-www.elra.info/Catalogue.html>, потоа одделот Text & Corpora на сајтот The linguistlist <http://linguistlist.org/sp/GetWRListings.cfm?WRAbbrev=Texts>. Корисна страница за студентите и наставниците по странски јазици е страницата на Католичкиот универзитет од Лувен, Белгија посветена на глотодидактичките корпуси на изучувачите на некој јазик: <http://www.uclouvain.be/en-cecl-lcworld.html>.

[3] Рускиот национален корпус е многу богат и во себе содржи повеќе видови корпуси: основен, синтаксички, медиски, паралелен, глотодидактички, дијалектолошки, поетски, говорен, акцентолошки, мултимедијален и историски.

[4] Кај нас во честа употреба е и терминот “компјутерска лингвистика” кој се користи во некои средини: германската, руската, полската, бугарската и други. Во голем дел други средини се користи англискиот термин (или пак негови калки) “computational”. Ние го претпочитаме францускиот термин “linguistique informatique” кој на прво место ја става улогата на информатиката, а не само на компјутерите, во лингвистичките истражувања.

[5] Многу богата со информации е страницата на каталонскиот лингвист Joaquim Llisterri посветена на говорните корпуси под наслов Speech and Spoken Language Resources <http://liceu.uab.es/~joaquim/language_resources/spoken_res/-biblio_corpus_orals.html>

збор е прикажан со масни букви. Инаку на тој сајт може да се вршат пребарувања во текстови од 230 еднојазични корпуси, меѓу кои и за македонскиот јазик за кого ексцерпцијата е вршена врз македонскиот дел од Wikipedia (<http://mk.wikipedia.org/wiki/>). Во однос на македонските корпуси треба да ги споменеме корпусите: Gralis (<http://www-gewi.uni-graz.at/gralis/>) (што го изработува Бранко Тошовиќ на Универзитетот во Грац, Австрија, и кој содржи и дел за македонскиот јазик) и Македонски електронски корпус [6], што го изработува Текстуалната лабораторија на Универзитетот во Осло, Норвешка (<http://www.tekstlab.uio.no/glossa/html/index_dev-.php?corpus=mak>).

Текстовите кои ги сочинуваат корпусите во најголем дел се специјално обработени или анотирани (морфолошки, синтаксички па и семантички) за да може да се пребаруваат и да се вршат разни операции врз нив. Токму така обработените текстови се од најголема полза за лингвистите. Тој начин на обработка на текстовите наоѓа најголема примена во електронските речници, било да се тоа онлајн (бесплатно или со претплата), со можност за симнување од интернет (бесплатно или со наплата) или пак на разни носачи (ЦД-Ром, ДВД-Ром и др.).

Денешните електронски речници претставуваат бази на податоци кои во себе содржат разни програми со чија помош корисниците може да дојдат до разни податоци содржани во речникот-база. Така, познатите речници како што се: Cambridge (<http://dictionary.cambridge.org/>), Collins Cobuild (<http://www.collinsdictionary.com/dictionary/english>), Chambers (<http://www.chambersharrap.co.uk/>), Longman (<http://www.ldoceonline.-com/>), Macmillan (<http://www.macmillandictionary.com/>), Merriam-Webster (<http://www.merriam-webster.com/>), Oxford (<http://oxforddictionaries.com/>) (за англискиот јазик), Le Trésor de la langue française informatisé (<http://atilf.atilf.fr/>), Le Robert (се претплатува), Antidote (се купува), Larousse (<http://www.larousse.fr/dictionnaires/francais-monolingue>), Linternaute (<http://www.linternaute.com/dictionnaire/fr/>) (за францускиот јазик), GRADIT (се купува), Zingarelli (<http://www.zanichelli.it/home/>) (се претплатува но можно е и гратис да се користи делумно онлајн), Garzanti (<http://www.garzantilinguistica.it/>) (треба да се отвори сметка), La Lessicografia della Crusca in Rete (<http://www.lessicografia.it/>), Sabatini-Coletti (<http://dizionari.corriere.it/dizionario_italiano/>), Gabrielli (<http://-

[6] За користење на корпусот потребно е да се отвори претходно сметка. Инаку до корпусот е овозможен пристап и преку сајтот на Џорџ Митревски кој е сместен на Универзитетот Оберн во Алабама, САД <http://www.auburn.edu/~mitrege/index.html>. Кај овој корпус може да се избере начинот на прикажување на резултатите да биде или преку одреден број на зборови или пак одреден број на реченици од левата и десната страна од бараниот збор.

www.grandidizionari.it/Dizionario_Italiano.aspx>), Devoto-Oli (се претплатува), Treccani (<http://www.treccani.it/vocabolario/>) (за италијанскиот јазик), DWDS (Das Digitale Wörterbuch der deutschen Sprache - <http://www.-dwds.de/>) DUDEN (<http://www.duden.de/>) (за германскиот јазик), DRAE (Diccionario de la lengua española de la Real Academia Española - <http://-rae.es/>) (за шпанскиот јазик) всушност претставуваат големи бази на податоци во кои може, особено во наплатната верзија, да се извршуваат огромен број на пребарувања кои се од суштинска важност за лингвистичките истражувања.

Можноста за пребарувања зависи од подготовката и обработката на лексикографскиот материјал. Доколку речничката граѓа е обработена преку база на податоци со доволен број на полиња кои ги предаваат суштинските одлики на јазичните единици, тогаш и програмите за пребарување ќе можат лесно и брзо да ги извлечат податоците од речникот. Во таа смисла најбогати, според наша оценка, се италијанскиот речник GRADIT (Grande **dizionario italiano** dell'uso) и францускиот речник TLFi (Trésor de la langue française informatisé). Преку нивно пребарување (првиот на ЦД-РОМ а вториот на интернет и на ЦД-РОМ), може да се добијат корисни податоци, преку ознаките што ги содржат (граматички, лексиколошки, семантички, социолингвистички, фразеолошки, паремиолошки). Во доменот на примерите што ги содржи, најбогат е оддалеку речникот Collins Cobuild (издание на ЦД-РОМ) кој за одредени зборови содржи и близу 19.000 примери.

Многу важни за разбирање и проверка на значењата на зборовите претставуваат визуелните речници кои ги придружуваат зборовите со цртежи во боја. Така, на платформата <http://www.ikonet.com> се наоѓаат визуелни речници за следните четири јазици: Visual Dictionary (за англиски, француски и шпански јазик) и Bildwörterbuch (<http://www.bildwoerter-buch.com/>) (за германски јазик). Терминот *визулен речник* се користи исто така и за специјално дијаграмско виуелно претствување на односите меѓу зборовите. Такви се следните речници: Visuwords (<http://www.visuwords.-com/fullscreen/>), Visual Thesaurus (<http://www.visualthesaurus.com/>) (за англиски, француски, германски, шпански, италијански и холандски јазик - со претплата), Vordvis (The visual dictionary) (<http://wordvis.com/>), Lexipedia (за англиски, француски, германски, шпански, италијански и холандски јазик) (http://www.lexipedia.com/).

Зборувајќи за важните лексикографски проекти, треба да ги споменеме проектите Wordnet 3.1 на Универзитетот во Принстон, САД (<http://wordnetweb.princeton.edu/perl/webwn>) и Wiktionary (<http://www.-wiktionary.org/>) кои обработуваат и други јазици освен англискиот.

Во однос на корисноста за лингвистот треба да го наведеме модулот MID (<http://mid.mozdev.org/installation.html>) што се додава како екстензија во прегледувачот Mozilla Firefox. Изборот на речниците, а ги има

повеќе од 2.000 онлајн еднојазични и двојазични речници, го врши самиот корисник и тој може да се сними за некоја подоцнежна употреба.

Во македонскиот виртуелен простор треба да се споменат следните речници: Makedonski.info (Дигитален речник на македонскиот јазик <http://www.makedonski.info/>), Rechnik.on.net.mk (<http://rechnik.on.net.mk/> - лексикон и правопис на македонскиот јазик и 10 двојазични речници со македонска паралела: англиски, германски, албански, грчки, италијански, француски, словенечки, турски, српски и руски), Македонско↔англиски речник, (кој се наоѓа на порталот Time.mk), Дигитален германско↔македонски речник (<http://makedonisch.info>) како и Поимник на зборови од областа на информатичката технологија (<http://www.mio.gov.mk/files/pdf/POIMNIK.pdf>).

Кога сме веќе кај корисноста на одредени опции, треба да спомeнеме дека покрај специјализираните програми на пример за синтеза на говор, денес постојат и одредени модули кои тоа можат да го направат во рамките на одредени општи програми. Така во прегледувачот Chrome на Гугл постои екстензијата Speakit! (<https://chrome.google.com/webstore/detail/speakit/pgeolalilifpodheeocdmbhehgnkkbak?hl=fr>) (што треба да се отвори во самиот Chrome и да се инсталира) која го претвора текстот што сме го маркирале на која и да интернетска страница во говор и тоа за повеќе разни јазици (text-to-speech).

Една друга корисна можност за пребарување на текстови напишани во форматот pdf се наоѓа во програмата Adobe Acrobat Reader Pro XI (се купува) која може во опцијата НАПРЕДНАТО ПРЕБАРУВАЊЕ да пребарува зборови или изрази во еден или повеќе текстови кои се наоѓаат на нашите дискови и потоа да ги сними резултатите во форма на конкорданци во посебен фајл што потоа можеме кога сакаме да го употребиме. Ова е одлична опција што ја нема ниту во професионалните програми за пребарување со кои можеме да вршиме пребарувања во нашиот компјутер и на него прикачените дискови.

Таквите програми го носат називот **десктоп пребарувачи** (desktop search engines) и се многу корисни за лингвистот: Google Desktop (<http://www.filehippo.com/fr/download_google_desktop/>) (кој веќе не се актуелизира од страна на Гугл), Copernic Desktop Search (<http://www.copernic.com/fr/products/desktop-search/>) (се купува верзијата Pro), DtSearch Desktop (се купува), File Locator Pro (се купува), DocFetcher (<http://docfetcher.sourceforge.net/en/download.html>) и др.

II. ЕЛЕКТРОНСКИ АЛАТКИ ЗА СОБИРАЊЕ ЛИНГВИСТИЧКИ ПОДАТОЦИ

Кога веќе еднаш лингвистот ќе ги собере јазичните податоци по електронски пат, тој треба да ги обработи според предметот и целта на

своето истражување, за да може да утврди одредени ставови или тези со кои ќа ја опише и/или објасни дадената појава.

Оваа втора група на податоци се всушност метаподатоци кои како свој предмет ги имаат јазичните податоци. Нив можеме да ги поделиме во две подгрупи: А. Општи лингвистички податоци и Б. Лингвистички податоци за определени јазични податоци.

А. Електронски алатки за собирање општи лингвистички податоци

Во оваа група на електронски алатки влегува сѐ она што може да му помогне на лингвистот да пронајде информации за лингвистиката, за сите нејзини гранки и за оние дисциплини кои се поврзани со лингвистиката. Оваа група е се разбира непресушна и секојдневно станува сѐ побогата, зашто постојано се создаваат нови сајтови, се креираат нови интернетски страници или се поставуваат бројни фајлови на статии, презентации, зборници, книги, терминолошки речници, енциклопедии и др. Наша цел во оваа етапа од излагањето е да претставиме само еден мал дел од можностите што ги нуди интернетот во оваа област.

И тука можеме да направиме одредено групирање и тоа во две категории: 1. пребарувачи (општи или специјализирани) на лингвистички документи и 2. архиви/депоа/репозиториуми на лингвистички документи (заеднички или лични). И тука, како и кај претходните алатки среќаваме сајтови кај кои се кумулираат и двете категории односно кои претставуваат истовремено пребарувачи и архиви на документи.

1. Пребарувачи (општи или специјализирани) на лингвистички документи

Ќе почнеме прво од пребарувачите кои во денешното доба на семантички веб почнуваат да нудат и поструктурирани резултати при пребарувањата. Така на пр. ако во Гугл го бараме зборот "linguistcs" во добиените резултати, на крајот од страницата, ќе видиме дека се јавува следниот текст: Searches related to **linguistics** linguistics **definition** linguistics **articles** linguistics **careers** linguistics **jobs** linguistics **terms** linguistics **major** linguistics **degree** linguistics **graduate programs**. Секој од тие поими е поврзан со зборот од нашето пребарување и, доколку кликнеме на секој од тие хиперлинкови ќе се отворат други резултати кои се поврзани со пребаруваниот збор. Од друга страна, речиси секој од добиените резултати содржи, од своја страна и хипероознака која може да нѝ отвори други страници поврзани со дадената страница. Таа ознака се содржи на крајот од вториот ред во кој се наоѓа електронската адреса на страницата и е обележена со еден рамностран триаголник свртен со рамната основа нагоре. Кога ќе кликнеме на триаголникот, се отвора мени со една, две или три опции: Cached, Similar и Share. Со кликнување на првата опција се отвора текстот кој е содржан во

меморијата на Гугл и таа опција е корисна во случај одредени страници да ги снема од интернет, што се случува. Ако притиснеме на вторaта опција, ќе се отворат нови страници кои се слични на страницата врз која се наоѓаме во тој момент.

Пребарувачот Yahoo ја содржи само ознаката Cached за секој резултат додека ги нема другите две опции од Гугл. Сугестиите за слични страници што се наоѓаат во дното од страницата кај Jahoo се следните: Also Try linguistics **jobs** linguistics **degree** **applied** linguistics **language** linguistics **introduction to** linguistics **definition of** linguistics **computational** linguistics linguistics **major**.

И пребарувачот Bing ја нуди само опцијата Cached page, додека сугестиите за слични страници за истиот поим "linguistics" се следните: Related searches for **linguistics Careers in** Linguistics **Jobs in** Linguistics **Introduction to** Linguistics Linguistics **Degrees** Linguistic **Studies** Linguistics **Books** Linguistic **Theory Anthropology**.

Во првата група треба уште да го споменеме и сервисот Scholar на Гугл (<http://scholar.google.com/>), кој претставува пребарувач на информации за дадена област и кој истовремено ги обележува оние електронски документи кои се наоѓаат на соодветните адреси во формат соодветен за симнување (pdf, doc). Овој сервис е пред сè погоден како средство за запознавање со насловите на трудовите кои третираат одредена проблематика, според тоа и за изготвување библиографија, а доколку сакаме да пронајдеме за одредена тема фајлови кои сакаме да ги симнеме тогаш е многу подобра опцијата во самиот Гугл за пребарување со помош на операторот filetype. Така на пр. во Гугл го пишуваме следното барање: **linguistics filetype:pdf** (или пак:doc, ppt), по што добиваме резултати само со фајлови во pdf или друг формат кои се однесуваат на бараниот збор и, се разбира, можеме да ги симнеме тие фајлови во нашиот компјутер. Истиот параметар **filetype:** може да се користи и во Yahoo, додека Bing не го содржи.

Во групата на специјализирани пребарувачи ќе ги споменеме најпрвин сајтовите на отворени електронски списанија: сервисот DOAJ (Directory of Open Acces Journals) (<http://www.doaj.org/>), кој содржи 581 списание од областа на лингвистиката и книжевната наука (консултирано на ден 30.11.2013). Сите тие списанија се онлајн и слободни се за користење, а статиите што се објавени во нив може да се симнат. На полето на Франкофонијата треба да го наведеме сервисот ISIDORE (<http://www.-rechercheisidore.fr/index>) кој пребарува и наоѓа документи спремни за симнување од хуманитарни и општествени науки од франкофонски сајтови. Во групата на специјализирани пребарувачи се издвојува сајтот LINGUIST List (<http://linguistlist.org/>), кој нуди разнородни интересни лингвистички податоци.

Лингвистот секогаш има потреба од добра дефиниција на лингвистичките термини. Во таа смисла корисни се следните сајтови на електронски лингвистички речници (енциклопедии): Systematic Dictionary of Corpus Linguistics (<http://donelaitis.vdu.lt/publikacijos/SDoCL.htm>), Glossary of linguistic terms (<http://www-01.sil.org/linguistics/GlossaryOfLinguisticTerms/>), Lexicon of Linguistics (<http://www2.let.uu.nl/UiL-OTS/Lexicon/>), ODLT (The Online Dictionary of Language Terminology - <http://www.odlt.org/>), A Glossary of Linguistic Terms (<http://www.cs.bham.ac.uk/~pxc/nlp/nlpgloss.html>), Lexikologie.de (The online-dictionary of technical terms in lexicology - <http://lexikologie.perce.de/index-en.php>), Electronic Glossary of Linguistic Terms - <http://www-01.sil.org/mexico/ling/glosario/E005ai-Glossary.htm>), Sémanticlopédie (Le dictionnaire des notions de sémantique utilisées en linguistique formelle - <http://www.semantique-gdr.net/dico/index.php/Accueil>), Enciclopedia Dell'italiano, (<http://www.treccani.it/enciclopedia/tag/dire/Enciclopedia_dell%E2%80%99Italiano/>), Словарь лингвистических терминов (<http://dic.academic.ru/dic.nsf/lingvistic/315>), Словарь лингвистических терминов (<http://www.textologia.ru/slovari/lingvisticheskie-terminy/?q=484> О. С. АХМАНОВА Словарь лингвистических терминов (<http://www.classes.ru/grammar/174.Akhmanova/source/worddocuments/_51.htm>).

Тука ќе споменеме и некои други сервиси слободни за користење. Тоа се најпрвин четирите франкофонски портали кои нудат стотици отворени електронски списанија: Persee.fr (<http://persee.fr/web/guest/home>) (содржи списанија од лингвистиката и од други хуманитарни и општествени науки), Revues.org (<http://www.revues.org/>) (хуманитарни и општествени науки), Erudit.org (<http://www.erudit.org/>) и Cairn.info (http://cairn.info/). Од овој белгиски сајт статиите може да се симнуваат слободно од сите компјутери приклучени на универзитетска мрежа на УКИМ, со оглед на тоа што нашиот универзитет има донација од франкофонската заедница во Белгија, додека од дома слободни за симнувања се единствено статиите објавени пред две или три години, за најголем дел од списанијата. Потоа имаме и еден шпански портал: Dialnet (<http://dialnet.unirioja.es/>) и нам поблиските сервиси: хрватскиот Hrčak (<http://hrcak.srce.hr/>) и српските Scindex (<http://scindeks.ceon.rs/>) и Kobson (<http://kobson.nb.rs/nauka_u_srbiji/casopisi_u_crossref-u_%28doi%29.22.html >).

Една од највредните работи за еден научник па според тоа и за лингвистот се докторските тези. Ќе наведеме шест адреси на кои може да се најдат и бројни тези од лингвистиката коишто може да се симнат. Тоа се: еден американски сајт - NDLTD (Networked Digital Libray of Theses and

Dissertations - <http://www.scirus.com/>[7]), еден европски сајт – DART-Europe (Digital Access to Research Theses - <http://www.dart-europe.eu/basic-search.php>), еден канадски сајт - Thesescanada (<http://collectionscanada.gc.-ca/thesescanada/>), два француски сајта – Thèses France (<http://www.-theses.fr/>) и TEL (Thèses en ligne - <http://tel.archives-ouvertes.fr/>) и еден италијански сајт Open Access in Italia (<http://wiki.openarchives.it/index.-php/Dati_e_cifre_sull%27Open_Access_in_Italia_-_2011>).

2. Архиви/депоа/репозиториуми на лингвистички документи (заеднички или лични)

Секако многу корисни за собирање на општи лингвистички податоци се архивите. Меѓу позначајните лингвистички архиви ќе ги наведеме: ACL Anthology (A Digital Archive of Research Papers in Computational Linguistics - <http://www.aclweb.org/anthology/>), Semanticsarchive.-net (<http://semanticsarchive.net/>), Hal articles en ligne (<http://hal.archives-ouvertes.fr/hal-00369078/fr/>), HAL – SHS (Hyper Article en ligne – Sciences de l'homme et de la société - <http://halshs.archives-ouvertes.fr/>), DSpace@-MIT (MIT's institutional repository - <http://dspace.mit.edu/>).

Многу корисни за лингвистот се и општите архиви кои содржат и огромен број на лингвистички книги. По својот голем опфат се издвојуваат најпрвин следните 4 архиви: Internet Archive (http://archive.org/index.php), Google books (<http://books.google.com/bkshp?hl=en&tab=wp>), Gallica (<http://gallica.bnf.fr/>) и Europeana (<http://europeana.eu/>).

Од непроценлива вредност се архивите на понови книги кои се засноваат на принципот на симнување Bit Torrent [8]: Uz-translations (<http:-//uz-translations.net/>), Rutracker (http://rutracker.org/forum/index.php), Libgen (<http://lib.free-college.org/>), Scribd (<http://fr.scribd.com/>), AvaxHome (<http://avaxhome.ws/>), Ebookee (<http://ebookee.org/>), BookFinder (<http:-//en.bookfi.org/>), Книги (<http://knigi.b111.org/>), Youscribe (<http://www.-youscribe.com/>), Ebook3000 (<http://www.ebook3000.com/>), Booktracker (<http://booktracker.org/>).

За француската лингвистика многу се корисни архивите на светскиот конгрес на француската лингвистика (досега се одржани 3 конгреси и сите ги има онлајн): CMLF 2008 (Congrès Mondial de Linguistique Française

[7] Инаку, сервисот Scirus на издавачката куќа Elsevier, кој ги има индексирано, покрај другото, и докторските дисертации од базата NDLTD, ќе згасне, како што се најавува на сајтот на Scirus, од почетокот на 2014 година.

[8] Во голем дел од нив треба претходно да се отвори сметка за да се појават адресите од кои може да се симнат бараните книги, а понекогаш книгите се во архивен формат за чие дезархивирање е потребно да се внесе лозинката што е дадена на истата страница.

2008 - <http://www.linguistiquefrancaise.org/index.php?option=com_toc-&url=/articles/cmlf/abs/2008/01/contents/contents.html>), CMLF 2010 (2ème Congrès Mondial de Linguistique Française - <http://www.linguistiquefrancaise.org/index.php?option=com_toc&url=/articles/cmlf/abs/2010/01/contents/contents.html>), CMLF 2012 (3e Congrès Mondial de Linguistique Française - <http://www.shs-conferences.org/index.php?option=com_toc&url=/articles/shsconf/abs/2012/01/contents/contents.html>).

Во однос на македонските архиви на лингвистички документи ќе ги споменеме неколкуте архиви со по некои лингвистички документи: Дигитална библиотека на Македонија (при НУБ "Св. Климент Охридски", Скопје - <http://www.dlib.mk>), Центар за ареална лингвистика при МАНУ (E-Publishing – 14 книги од лингвистика - < http://ical.manu.edu.mk/index.php/publications>), Дигитален архив на македонскиот јазик (13 лингвистички текстови - <http://damj.manu.edu.mk/>), Библиографија на македонскиот јазик (<http://bmj.manu.edu.mk/>), Меѓународен семинар за македонски јазик, литература и култура при УКИМ (предавања и Научна конференција (за лингвистика и за литература, со 7 годишта – 2004, 2005, 2006, 2007, 2008, 2010, 2011 - < http://www.ukim.edu.mk/mk_content.php?meni=96-&glavno=34>), Институт за македонски јазик "Крсте Мисирков", (адресата за симнување на одредени броеви од списанието Македонски јазик (Македонски јазик, год. LXII, 2011, 221 стр., Македонски јазик, год. LXI, 2010, 266 стр., Македонски јазик, год. LX, 2009, 370 стр., Библиографија на македонски јазик (1950-2009), LX, 100 стр.) е следната: http://imj.ukim.edu.mk/index.php?option=com_content&view=article&id=106&Itemid=18, додека адресата за симнување на еден број од списанието Македонистика (Македонистика бр. 10, Скопје, 2008, 247 стр.) е следната: <http://imj.ukim.edu.mk/images/stories/makedonistika.pdf>), Philological Studies (<http://philologicalstudies.org/>), Репозиториј на Универзитетот "Гоце Делчев" – Штип (за јазик и литература - <http://eprints.ugd.edu.mk/view/subjects/LL.html>), Годишен зборник на Филолошкиот факултет при Универзитетот "Гоце Делчев" – Штип (<http://e-lib.ugd.edu.mk/resursi/zbornici/filoloski/-Zbornik_Filoloski_2011.pdf>)

Завршувајќи го овој дел, ќе споменеме дека постојат голем број на заеднички архиви на разни лингвистички институти во разни земји. Во секој од тие архиви се содржат трудови на членовите на институтите или на истражувачките екипи или на лингвистите наставници на универзитетите. Такви се на пример страниците: William Labov (<http://www.ling.upenn.edu/~wlabov/home.html>), Jacques Moeschler <http://www.unige.ch/lettres/linguistique/moeschler/enseignements/cours.php>), Marko Tadić (<http://www.hnk.ffzg.hr/mt/>). Но постојат и веб страници на лингвисти што ќе ги наречеме слободни стрелци кои си имаат направено лични веб страници/сајтови. Тука ќе ги споменеме страниците на следните лингвисти: David Crystal (<http://www.davidcrystal.com/>), Trad.it, il sito di

Bruno Osimo (<http://www.trad.it/>), Patcrick Charaudeau (<http://www.patrick-charaudeau.com/>) Robert de Beaugrande (<http://www.beaugrande.-com/>), George Lakoff (<http://georgelakoff.com/>).

Б. Електронски алатки за собирање лингвистички податоци за определени јазични податоци

Веќе споменавме погоре дека денес во електронските бази на податоци и во електронските речници и двата вида на податоци, јазичните и лингвистичките, може да се најдат честопати на едно место. Така на пр. ако ги проучуваме јазичните единици кои му припаѓаат на разговорниот/-колоквијален регистар, ние ќе треба да ги пронаоѓаме тие единици самите, во говорната интеркомуникација или во пишаните дела, но освен тоа треба да им определиме и бројни други карактеристики или ознаки со кои тие ќе можат да бидат лингвистички опишани или објаснети: нивната форма (звучна и пишана), нивните морфолошки, синтаксички, семантички, прагматички, стилистички и социолектни карактеристики. Како што видовме погоре, еден дел од електронските речници, оние најдобрите, содржат голем дел од тие карактеристики. Така речникот на италијанскиот јазик GRADIT овозможува да се пребарува речничката граѓа според дваесетина критериуми. На тој начин таквите речници претставуваат вистински бази на податоци за одредени јазични параметри и можат донекаде да ја заменат потребата од создавање на прави бази на податоци.

Базите на податоци се неопходни за едно сериозно лингвистичко истражување. Меѓу бројните програми за креирање и управување со бази на податоци секако дека најпристапна е програмата Microsoft Access која може да биде доволна доколку не се работи за огромен број на записи. Лингвистичкото истражување кое се заснова на база на податоци и кое е добро изведено може да ги понесе епитетите за коишто стана збор во почетокот на ова излагање: **побрзо, поопфатно, попрецизно, попроверливо, поевтино** и **попристапно**.

Тоа ќе биде *побрзо* зашто може за неспоредливо пократко време да обработи многу поголем број на податоци отколку класичното лингвистичко истражување.

Тоа ќе биде *поопфатно* зашто се заснова на огромен број на електронски документи репрезентативни за дадениот јазик и затоа што на една база можат истовремено да работат повеќе луѓе што не е можно при класичното лингвистичко истражување.

Тоа ќе биде *попрецизно* зашто овозможува да се дојде до детални информации за многу јазични појави за кои инаку не би можело да се дојде преку класичното лингвистичко истражување.

Тоа ќе биде *попреверливо* затоа што базата и резултатите што ќе произлезат од неа во едно такво истражување може да се проверат во секое време од други истражувачи.

Тоа ќе биде *поевтино* затоа што еднаш создадените бази остануваат засекогаш отворени и на достап на лингвистите кои не треба повторно да трошат средства за истото истражување и заштo може постојано да се надоградуваат со нови записи.

Тоа ќе биде *попристапно* затоа што базата по електронски пат може да им биде достапна на огромен број на луѓе и затоа што резултатите од истражувањето претставени во отворена електронска форма можат да ги користат многу поголем број луѓе отколку при класичното лингвистичко истражување.

БИБЛИОГРАФИЈА

BAKER, Paul & HARDIE, Andrew & MCENERY, Tony: *A Glossary of Corpus Linguistics*, Edinburgh University Press, 2006, 187 p.

BOLSHAKOV, Igor A. & GELBUKH, Alexander: *Computational linguistics. Models, Resources, Application*s, Instituto Politécnico Nacional, **Mexico City,** 2004 186 p. <http://www.gelbukh.com/clbook/>

CHIARI, Isabella: *Introduzione alla linguistica computazionale*, Editori Laterza, Bari, 2007, 210 p.

HABERT, Benoît: Instruments et ressources électroniques pour le français**,** Ophrys (“L'essentiel français**”),** Gap/Paris, 2005, 176 p.

HUNSTON, Suzan: *Corpora in Applied Linguistics*, Cambridge University Press, 2002, 254 p.

LAWLER, John M. & DRY, Helen Aristar (eds.): *Using Computers in Linguistics. A Practical Guide,* Routledge, London/New York, 2003 (1998), 297 p.

LÜDELING, Anke & KYTÖ, Merja (eds.): *Corpus Linguistics - An International Handbook* Vol. 1, Walter de Gruyter, Berlin-New York, 2008, pp. 1-776.

LÜDELING, Anke & KYTÖ, Merja (eds.): *Corpus Linguistics - An International Handbook* Vol. 2, Walter de Gruyter, Berlin-New York, 2009, pp. 777-1353.

NESSELHAUF, Nadja: *Collocations in a Learner Corpus*, John Benjamins Publishing Company, Amsterdam/Philadelphia, 2005, 331 p.

OOI, Vincent B.Y.: *Computer Corpus Lexicography*, Edinburgh Universit Press, 1998, 258 p.

Proceedings of the Corpus Linguistics Conference, University of Liverpool, UK, 2009. <http://ucrel.lancs.ac.uk/publications/cl2009/>

SINKLER, John : *Reading Concordances, An Introduction,* Pearson Education Limited, London, 2003, 180 p.